Abstract Title Page
Not included in page count.

Title: Multilevel Propensity Score Matching Within and Across Schools

Author(s): Benjamin Kelcey (ben.kelcey@gmail.com)

Abstract Body

Limit 5 pages single spaced.

Background / Context:

A central issue in nonexperimental studies is the identification of comparable individuals (e.g. students) to remove selection bias. One such increasingly common method to identify comparable individuals and address selection bias is the propensity score (PS). PS methods rely on a model of the treatment assignment to identify comparable individuals on the basis of similar probabilities of receiving treatment. To extend PS methods to education data, the methods must be appropriately adapted to address the potential multilevel nature of treatment assignments. For instance, in studies that resemble multi-site randomized trials, treatments are assigned to students within each school. Such assignments are often based on student and school characteristics while the mechanism of assignment may vary substantially across schools. Correspondingly, selection bias may originate from both student and school levels. One recent examination of this issue includes estimating the effect of kindergarten retention on achievement where students' probability of being retained is thought to be a function of the student's personal characteristics but also his/her school membership (Hong & Raudenbush, 2006).

To attend to how the treatment selection mechanism differs among schools, two principal approaches have been proposed. The first approach makes use of a canonical single level PS in conjunction with restricting comparisons to individuals within the same school (Rosenbaum, 1986). This approach addresses the varying influence of a school by blocking on school membership, and thus on observed and unobserved school covariates. A second approach to address varying school influence on treatment assignment is use of multilevel PSs (Hong & Raudenbush, 2006; Kim & Seltzer, 2007). This approach explicitly models how treatment assignment mechanisms differ among schools potentially making PSs comparable across school boundaries. In using such approaches to identify comparable students and estimate effects, both approaches hold attractive features and potential drawbacks. In particular, use of single level PSs which constrain comparisons to students within schools may alleviate or protect against residual bias from unmeasured school level factors that influence the treatment assignment. However, such protection is often coupled with a lack of comparable matches within each school. In particular, restricting comparisons to within schools generally requires a large reservoir of treatment and control students in each school to ensure high quality matches. In contrast, by taking into account both school membership and how the predictive capacities of covariates vary across schools, multilevel PSs can more effectively match students across schools. That is, because multilevel PSs take into account the extent to which school influence on student treatment assignment varies, estimated probabilities of receiving treatment are more comparable across schools. As a result, the reservoir of potential matches is much larger and often leads to higher quality matches. Such advantages, however, are also paired with the assumption that there are no unmeasured school level covariates. Because multilevel PSs permit matches across schools, they assume adequate adjustment for the differences in the selection processes.

Purpose / Objective / Research Question / Focus of Study:

In this study, we evaluated the performance of both of these approaches in nonexperimental studies resembling multisite randomized trials. In particular, we assessed the performance of matching students across schools based on multilevel PSs and matching students within schools based on single level PSs with specific focus on the sensitivity of the PSs to omitted covariates. More specifically, we examined the extent to which matching across groups

using a multilevel PS produces a more effective estimator of the treatment effect as compared to matching within groups using a single level PS. We addressed the research question by assessing the performance of the treatment effect estimators in multiple scenarios using Monte Carlo simulation. For proposal brevity, we highlight four simple situations.

1. No unmeasured covariates that influence the treatment assignment
2. One unmeasured school level covariate that influences the treatment assignment
3. One unmeasured student level covariate that influences the treatment assignment
4. One unmeasured school level covariate and one unmeasured student level covariate; both of which influence the treatment assignment

Setting:
(May not be applicable for Methods submissions)

Population / Participants / Subjects:
(May not be applicable for Methods submissions)

Intervention / Program / Practice:
(May not be applicable for Methods submissions)

Significance / Novelty of study:

Multilevel PSs potentially hold means to both remove selection bias associated with how selection processes differ among schools as well as from bias associated with omitted school level variables. Accordingly, multilevel PSs estimate students' probabilities of receiving treatment in a way that is both robust and more comparable across schools. In turn, such features frequently promote higher quality matches and tend to remove more selection bias. Although matching within schools lays claim to high face validity, the potential bias reduction gained by allowing across school matches has important implications. In particular, studies which examine treatments of low incidence (e.g. retention) are well positioned to gain from across school matching based on multilevel PSs. Similarly, matching across schools based on multilevel PSs in studies which examine treatments implemented at the teacher/classroom level may improve estimation. That is, because evaluating treatments implemented at the teacher/classroom level requires matching intact classrooms, the potential matches within a school (e.g. another classroom in that school and grade) can be very limited. In such situations, a method which affords researchers the ability to identify comparable classrooms in other schools may be highly beneficial and allow evaluation of treatments that otherwise could not have been evaluated.

Statistical, Measurement, or Econometric Model:

We performed four Monte Carlo simulations, each with 1,000 simulated data sets. The first simulation examined the performance of matching within and across schools based on both multilevel and single level PSs. The remaining simulations examined the approaches sensitivity to omitted variables at both the student and school levels. We generated the individual characteristics (X_1,X_2,X_3) and group characteristics (W_1,W_2,W_3) using independent normal distributions with mean 0 and variance 1. Sample sizes presented here focus on 100 schools with an average of 10 students per school. We designed the treatment as the realization of a dichotomous variable given student characteristics and school characteristics. For each simulation, the true treatment assignment mechanism followed the hierarchical generalized linear model:

$$
\begin{aligned}
Level\,1: logit(P(Z=1)) &= \beta_0 + \sum_{m=1}^{3} \beta_m X_{mij} \\
Level\,2: \beta_0 &= \gamma_{00} + \sum_{n=1}^{3} \gamma_{n0} W_{nj} + u_{0j} \\
\beta_m &= \gamma_{0m} + \sum_{n=1}^{3} \gamma_{nm} W_{nj} + u_{mj}
\end{aligned}
\tag{0.1}
$$

for random effects u~multivariate normal with mean 0 and covariance matrix

$$
\sum u_{treatment} = \begin{pmatrix} 0.125 & 0.035 & 0.035 & 0.035 \\ 0.035 & 0.125 & 0.035 & 0.035 \\ 0.035 & 0.035 & 0.125 & 0.035 \\ 0.035 & 0.035 & 0.035 & 0.125 \end{pmatrix} \tag{0.2}
$$

The true outcome model for each simulation was a random intercept and slope hierarchical linear model (HLM)

$$
\begin{aligned}
Level\,1: Y_{ij} &= \beta_0 + \delta Z_{ij} + \sum_{m=1}^{3} \beta_m X_{mij} + \varepsilon_{ij} \\
Level\,2: \beta_0 &= \gamma_{00} + \sum_{n=1}^{3} \gamma_{n0} W_{nj} + u_{0j} \\
\beta_m &= \gamma_{0m} + \sum_{n=1}^{3} \gamma_{nm} W_{nj} + u_{mj}
\end{aligned}
\tag{0.3}
$$

where δ was a treatment effect of 0.5, Z is the treatment assignment, $\varepsilon \sim N(0, \sigma^2_{outcome})$ and u_0 $\sim MVN(0, \tau_{outcome})$

$$
\tau_{outcome} = \begin{pmatrix} 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 \end{pmatrix} \tag{0.4}
$$

For proposal brevity we limit our discussion to simulations in which each coefficient in (0.1) equals ±0.3 whereas the coefficients in (0.3) equal ±0.5.

To estimate each student's probability of receiving the treatment we constructed PSs using different versions of (0.1) and highlight six different PS specifications which constrain various parameters in (0.1). The specifications are:

(1) Constrain all random effects (u) and cross level interactions (γ_{nm}) to be zero (*Single level main fixed effects only*)

(2) Constrain all random effects (u) (*Single level all fixed effects*)

(3) Constrain slope random effects (u_{mj}) and cross level interactions (γ_{nm}) to be zero (*Multilevel with a random intercept and main fixed effects only*)

(4) Constrain slope random effects (u_{mj}) to be zero (*Multilevel with a random intercept and all fixed effects*)

(5) Constrain both school fixed effects (γ_{n0}) and cross level interactions (γ_{nm}) to be zero (*Multilevel with random intercept and slopes and only level one main fixed effects*)

(6) Do not constrain any parameters in (0.1) (*Multilevel with random intercept and slopes and all fixed effects*)

To estimate the treatment effect for each data set and approach, we combined adjustment on the PS with a HLM (e.g. Hong & Raudenbush, 2006; e.g. Hirano & Imbens, 2002; Kleyman & Hansen, 2008). Here, we limit our discussion to matching students within or across schools using full matching so as to make use of all available students (e.g. Hansen, 2004). Using match indicators, we modeled the outcome as a HLM as follows:

$$\begin{aligned} Level\,1: Y_{ij} &= \beta_0 + \hat{\delta} Z_{ij} + \sum_{q=1}^{Q} \beta_q M_{qij} + \varepsilon_{ij} \\ Level\,2: \beta_0 &= \gamma_{00} + u_{0j} \end{aligned} \tag{0.5}$$

where $\hat{\delta}$ is the estimated treatment, M is the matched group, and Q is the number of matched groups minus one. To compare the estimates based on the different classes and specifications of models, we used the results of the Monte Carlo simulations to estimate the bias and mean-squared error (MSE) of each approach. We estimated these quantities using:

$$\widehat{Bias} = \frac{1}{N}\sum_{i=1}^{N} \hat{\delta}_i - \delta \tag{0.6}$$

and

$$\widehat{MSE} = \frac{1}{N}\sum_{i=1}^{N} (\hat{\delta}_i - \delta)^2 \tag{0.7}$$

where N represents the number of simulated data sets.

Research Design:
(May not be applicable for Methods submissions)

Data Collection and Analysis:
(May not be applicable for Methods submissions)

Findings / Results:

In the first simulation we examined the quality of treatment effect estimates when all covariates were observed. The results suggested that estimators which allow matching students across schools outperform estimators that restrict matches to within schools (Table 1). This result was anticipated as the main utility of restricting matches to within schools to eliminate unobserved school level bias was irrelevant. As a result, matching across school boundaries

tended to promote higher quality matches. Results from the first simulation additionally suggested that matches based on multilevel PSs tend to outperform estimators which match students based on single level PSs regardless of whether matches were within or across schools (Table 1 & Tables not presented here). The second simulation compared the performance of the different approaches when a school level confounding variable was omitted from the analyses. Evident from Table 2, the performance of schemes which matched across schools degraded whereas within schools matching schemes did not (Table 2). However, those approaches which matched across schools based on a multilevel PS still appreciably outperformed those approaches which matched within schools based on a single level PS (Table 2). Similar results held for scenarios which omitted a student and/or a school level covariate (Tables 3 & 4).

Usefulness / Applicability of Method:

The results presented primarily suggest that even with omitted variables there is an appreciable advantage to matching across schools when one employs multilevel PSs. To a large extent, PSs which use methods to address how selection processes differ among schools (e.g. multilevel PSs) tended to be robust to school level omitted confounding variables. That is, because multilevel PSs employ random effects, they can both summarize omitted school level variables and account for the across school variation in the predictive capacity of covariates. As a result, multilevel PSs potentially hold means to both remove selection bias associated with how selection processes differ among schools as well as from bias associated with omitted school level variables.

Conclusions:

In summary, the results of this study suggest that matching across schools based on multilevel PSs has much to offer and that further study into its utility and the conditions under which it outperforms matching within schools based on a single level PS need to be explicated. In particular, the simulations presented address small covariate spaces with simple variable relations. Study of designs with more complex associations among both unobserved and observed confounders is warranted. Finally, omitted in this summary are more complex details concerning covariate balance. Specifically, obtaining and assessing covariate balance both within and across schools to build a strong basis for inference is much more complex.

Appendices

Not included in page count.

Appendix A. References

Brookhart, M., Schneeweiss, S., Rothman, K., Glynn, R., Avorn, J., & Sturmer, T., (2006). Variable selection for PS models. *Practice of Epidemiology*, vol. 163, 12, pp. 1149-1156.

Hansen, B.B. (2004), 'Full Matching in an Observational Study of Coaching for the SAT', *Journal of the American Statistical Association*, **99**, 609–618.

Hirano, K., & Imbens, G., (2002). Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. Health Services & Outcomes Research Methodology, 2, pp. 259–278.

Holland, P., (1986). Statistics and Causal Inference. *Journal of the American Statistical Associations*, Vol. 81, No.396, p. 947.

Hong, G., & Raudenbush, S., (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data, *Journal of the American Statistical Association*, vol. 101, 475, pp. 901-910.

Kim, J., & Seltzer, M., (2007). Causal inference in multilevel settings in which selection processes vary across schools. *National Center for Research on Evaluation, Standards and Student Testing (CRESST)*, CSE Technical Report 708, University of California, Los Angeles

Kleyman, Y., & Hansen, B., (2008). Test for Covariate Balance in Comparative Studies. Paper presented at the Joint Statistical Meetings 2008 annual conference.

Rosenbaum, P., (1986). Dropping out of high school in the United States: An observational study. Journal of Educational Statistics, vol. 11, 3, pp. 207-224.

Appendix B. Tables and Figures

Table 1: Comparison of within school and across school matching using several different single and multilevel propensity scores approaches and assuming there are no omitted covariates

Approach	(1)	(1)	(2)	(2)	(3)	(4)	(5)	(6)
Matching	*Within*	*Across*	*Within*	*Across*	*Across*	*Across*	*Across*	*Across*
PS Type	*Single*	*Single*	*Single*	*Single*	*Multilevel*	*Multilevel*	*Multilevel*	*Multilevel*
Fixed Effects	*Main*	*Main*	*All*	*All*	*Main*	*All*	*Level 1 only*	*All*
Random Effects	*None*	*None*	*None*	*None*	*Intercept*	*Intercept*	*Intercept & Slope*	*Intercept & Slope*
10 x Bias	6.63	5.06	6.37	0.89	5.21	-0.18	-0.56	-0.53
10 x Var	0.37	0.40	0.36	0.28	0.40	0.29	0.28	0.29
10 x MSE	4.77	2.96	4.42	0.35	3.12	0.29	0.31	0.32

Note: We also examined, for instance, matching within schools based on multilevel PSs but do not present the results due to space limitations

Table 2: Comparison of within school and across school matching using several different single and multilevel propensity scores approaches and assuming there is one school level covariate that is omitted

Approach *Matching PS Type Fixed Effects*[a] *Random Effects*	(1) *Within Single Main None*	(1) *Across Single Main None*	(2) *Within Single All None*	(2) *Across Single All None*	(3) *Across Multilevel Main Intercept*	(4) *Across Multilevel All Intercept*	(5) *Across Multilevel Level 1 only Intercept & Slope*	(6) *Across Multilevel All Intercept & Slope*
10 x Bias	6.57	6.70	6.65	3.42	6.43	2.95	-0.52	-1.10
10 x Var	0.38	0.43	0.39	0.35	0.44	0.36	0.31	0.31
10 x MSE	4.70	4.92	4.80	1.52	4.57	1.23	0.34	0.43

[a]All fixed effects in Table 2 exclude W_3 a school level covariate which predicts the treatment assignment

Table 3: Comparison of within school and across school matching using several different single and multilevel propensity scores approaches and assuming there is one student level covariate that is omitted

Approach *Matching PS Type Fixed Effects*[b] *Random Effects*	(1) *Within Single Main None*	(1) *Across Single Main None*	(2) *Within Single All None*	(2) *Across Single All None*	(3) *Across Multilevel Main Intercept*	(4) *Across Multilevel All Intercept*	(5) *Across Multilevel Level 1 only Intercept & Slope*	(6) *Across Multilevel All Intercept & Slope*
10 x Bias	7.82	6.75	7.68	3.85	6.84	3.52	3.76	2.33
10 x Var	0.41	0.44	0.41	0.35	0.45	0.35	0.34	0.36
10 x MSE	6.52	4.99	6.31	1.82	5.13	1.59	1.76	0.90

[b]All fixed effects in Table 3 exclude X_3 a student level covariate which predicts the treatment assignment

Table 4: Comparison of within school and across school matching using several different single and multilevel propensity scores approaches and assuming there is one student level and one school level covariate that is omitted

Approach	(1)	(1)	(2)	(2)	(3)	(4)	(5)	(6)
Matching	*Within*	*Across*	*Within*	*Across*	*Across*	*Across*	*Across*	*Across*
PS Type	*Single*	*Single*	*Single*	*Single*	*Multilevel*	*Multilevel*	*Multilevel*	*Multilevel*
Fixed Effects[c]	*Main*	*Main*	*All*	*All*	*Main*	*All*	*Level 1 only*	*All*
Random Effects	*None*	*None*	*None*	*None*	*Intercept*	*Intercept*	*Intercept & Slope*	*Intercept & Slope*
10 x Bias	7.99	8.03	7.98	6.11	7.77	5.60	3.88	3.25
10 x Var	0.41	0.45	0.40	0.37	0.45	0.38	0.37	0.35
10 x MSE	6.80	6.89	6.77	4.10	6.48	3.51	1.88	1.41

[c]All fixed effects in Table 4 exclude X_3 a student level covariate and W_3 a school level covariate; both of which predict the treatment assignment